Lite-GISTexter at DUC2004

Finley Lacatusu, Andrew Hickl, Sanda Harabagiu and Luke Nezda
Language Computer Corporation
1701 N. Collins Blvd.
Richardson, TX 75080
{finley, andy, sanda, luke}@languagecomputer.com

Abstract

This paper discusses the document summarization techniques implemented in Language Computer Corporation's Lite-GISTexter for the DUC 2004 evaluation. We show that work in the summarization of single documents (as well as collections of documents) can benefit from conceptually-simple algorithms that make use of linguistic features of the texts themselves. Despite using only minimal lexico-semantic resources, Lite-GISTexter produced coherent and accurate summaries that received consistently high rankings across the tasks evaluated in DUC 2004.

1 Introduction

In this paper, we suggest that effective and accurate automatic document summarization can be achieved without resorting to complex algorithms or evaluation metrics. Although we believe that the best summarization systems will stem from a thorough understanding of discourse structure, we feel that until discourse structure information can be obtained automatically, work in summarization can also benefit from reliable algorithms that are conceptually-simple and linguistically-grounded.

Three different kinds of summaries were evaluated in DUC 2004: (1) very short single document summaries, (2) short multi-document summaries, and (3) short summaries generated in response to a request for biographical information about a named entity. Tasks 1 and 3 asked participants to create very short single document summaries no longer than 75 bytes in length. We assume that these kinds of very short summaries can be generated by identifying the most useful sentence in a document, and then by performing compression (if necessary) in order to reduce the sentence's length to the word- or byte-length imposed by the DUC organizers. Returning to a simplifying assumption first proposed in DUC 2003, we hypothesize that the first sentence in a document is often the most informative sentence in a document, and therefore, provides the best input for the generation of these headline-style summaries.

In Tasks 2 and 4, participants generated short summaries (of no longer than 665 bytes) for clusters of documents that shared a common topic or domain of interest. For these tasks we used topic signature terms to both identify information as relevant to a summary and to eliminate sources of redundancy from the output. Although topic signature-based approaches to single- and multi-document summarization have been presented at previous DUC competitions (Lin and Hovy, 2002), our system achieves competitive results for the short summarization task by ranking sentences and eliminating redundancy solely based on the co-occurrence of topic signature terms; previous systems have had to rely on ranking systems developed for QA or similarity measures such as Maximal Marginal Relevance (Carbonell and Goldstein, 1998) in order to obtain comparable results. After initially ranking sentences based on the topic signature method first described by Lin and Hovy (2000), we gauge whether a candidate sentence should be added to a summary based on the number of topic signature terms that occur in both the summary and the candidate sentence. If the number of overlapping terms is above a certain threshold, the candidate sentence is rejected; otherwise, the sentence is added to the summary.

In Task 5, participants were asked to create short summaries that provided answers to definition questions of the form *Who is X?*. Our system reliably identified the passages relevant for answering these definition questions by returning only those sentences that contain the name of the person of interest in an appositive and/or copula construction. The success of this simplifying assumption can be derived from the fact that news articles often introduce or identify named entities with short descriptions that involve these kinds of constructions. This approach garnered the best scores in responsiveness and coverage for this task (as scored by humans).

The rest of this paper describes the performance of our system, known as Lite-GISTexter, in the DUC 2004 evaluation. Section 2 describes our current approach to generating short summaries. Section 3 details our approach to generating very short summaries. Section 4 discusses our approach to generating short summaries focused by biographical questions. Finally, section 5 provides results from the DUC 2004 evaluation, and section 6 provides our conclusions and identifies some challenges we hope to address in future evaluations.

2 Generating short summaries

The creation of multi-document summaries depends on the extraction of a set of sentences whose content best describes the information content of the document collection as a whole. In order to identify the most relevant sentences to include in a summary, multi-document summarization systems must first construct a representation of the topic of a document collection. This topic representation should both approximate the conceptual content of the documents themselves and provide for the identification and extraction of a set of terms (or keywords) that are most likely to be associated with a particular topic. Once this topic representation is in place, sentence selection can proceed by identifying those sentences which sufficiently approximate the content of the topic itself.

Task 2 in DUC 2004 required participants to generate short summaries for each of 50 document clusters taken from the AP and New York Times newswires (1998-2000). In Task 4, participants generated similar types of short summaries from clusters of documents taken from the Arabic *Agence France Presse* newswire. Documents for this task were translated from Arabic into English by both human translators and machine translation systems. Unlike previous evaluations, however, participants in these tasks did not have access to prepared topic descriptions for each of the document sets.

Without a topic description to inform sentence selection for multi-document summarization, we decided to implement the methodology for automatically acquiring topic signatures first proposed in (Lin and Hovy, 2000). Developed originally for single-document summarization, Lin and Hovy's algorithm computes a weight for each term (excluding stop words) in a document cluster based on its relative frequency in a relevant set of documents. Following Lin and Hovy, we refer to the set of the highest ranked terms as comprising the *topic signature* for a document cluster.

(We also experimented with other various weighting schemes – including document cluster frequency and *tf*idf* – in order to detect topic signature terms, but the summaries produced using the Lin and Hovy algorithm consistently generated higher ROUGE scores than any other method.)

Instead of stemming terms, we use the *morphstr* function from WordNet in order to detect morphological variants of the same term. We also attempted to stem words using the Porter stemmer (Porter, 1980). However, resulting summaries produced from DUC 2003 data had consistently lower ROUGE scores than data stemmed using *morphstr*.

The calculation of weights for topic signatures terms depends on the classification of documents into sets of relevant and nonrelevant documents. For each topic cluster, we considered all of the documents in a topic cluster as relevant; the documents in the remaining TDT document clusters were taken as the nonrelevant set.

Each term in the cluster of documents to be summarized receives a score equal to the $-2log\lambda$ value.
$-2log\lambda = -2log\frac{L(H_1)}{L(H_2)}$, where:
$L(H_1) = b(O_{11}; O_{11} + O_{12}, p)b(O_{21}; O_{21} + O_{22}, p)$
$L(H_2) = b(O_{11}; O_{11} + O_{12}, p_1)b(O_{21}; O_{21} + O_{22}, p_2)$
O_{11} is the frequency of the term in the cluster
O_{12} is the frequency of the term in the other clusters
O_{21} is the frequency of the other terms in the cluster
O_{22} is the frequency of the other terms in the other clusters
p is the probability that the cluster is relevant to the topic, given the term, in the hypothesis H_1 - the relevancy of a document is independent of the term
p_1 is the probability that the cluster is relevant to the topic, given the term, in the hypothesis H_2 - the presence of the term indicates strong relevancy
p_2 is the probability that the cluster is relevant given the other terms, in hypothesis H_2

Since our summarization system depends on extracting whole sentences from a collection of documents, we used the computed weights to assign each sentence a composite score equal to the sum of the weights of all of its terms. All of the sentences in the relevant set of documents are then ranked based on this composite score.

Short summaries were then created according to the algorithm presented below. The first step involves adding the highest ranked sentence to the summary. As long as the summary is under the maximum length prescribed by the DUC evaluation (665 bytes), the next highest ranked sentence gets added to the summary if the number of topic signature terms it adds to the summary is greater than the number of topic signature terms that it has in common with sentences already found in the summary. For example, a sentence that contains 7 topic signature terms will be added to a summary if it has no more than 3 terms in common with the summary. In this way, the information content of a sentence relative to the summary can be measured by the number of terms from the topic signature that they both have in common. If the sentence contains no new topic signature terms – or if it contains fewer new terms than redundant terms – it is discarded, and the next highest ranked sentence is then considered.

Once the summary reaches 665 bytes (or greater) in length, sentence compression proceeds to reduce the overall length of the summary. If the adding of a sentence to the summary would create an overall length of more than 600 bytes, the sentence is not added to the summary and no more sentences are considered. If the summary is still too long, Lite-GISTexter truncated the summary at the last word prior to the 665 byte limit.

Step 1: Put the highest ranked sentence in the summary
Step 2: **If** $length(summary) > 665$ *then GOTO Step 4*
Else *add the next sentence if it adds more information to the summary, than it has in common with it. The information content of a sentence in our case is measured by the number of terms from the topic signature that appear in the sentence. For example if the sentence has 7 topic signature terms, and 3 of them appear also in the summary, then add it to the summary. If 6 terms from the sentence appear also in the summary, then disregard the sentence.*
Step 3: Repeat step 2 until we run out of space in the summary or candidate sentences
Step 4: Compress the summary
- **If** $length(summary - lastsentence) > 600$ *remove the last sentence*
- **Else** *for each summary sentence compress it as in Task 1, but with a much weaker grammar.*
- **If** *the result is still too large, cut the summary at the last word such that it is shorter than 665 Bytes.*

For Task 2, we prepared three different sets of summaries evaluation. The first set were generated by following the entire algorithm presented above. The second and the third set differed in terms of how much inter-sentence compression was performed. In the second set, sentences were no longer considered once the summary's length exceeded 600 bytes; in the third set, summaries were automatically truncated after the word immediately preceding the 665 byte mark. Although this final approach seems the least elegant of the three, it received the second-highest (ROUGE) score of all the systems. This result reinforces our belief that several of the highest-ranked sentences (as identified by the topic signatures method) can be considered important to the overall perceived quality of a summary. In these cases, it appears that less than 65 bytes of text can make a large difference in score.

We applied the same summarization system to Task 4 that we applied to Task 2. Since our summarization system depends on a statistical approach, we were able to successfully summarize the noisy data produced by machine translation systems just as well as the "smoother" data produced by human translators. See Table 1 for representative examples of summaries generated from manual and automatic translations.

3 Generating very short summaries

Very short document summaries enable a user to make an informed decision about the relative utility of the content of a document with minimal effort. Akin to document headlines, the creation of very short summaries require summarization systems to extract the single most important piece of information in a document and to represent it using as few words as possible.

Tasks 1 and 3 in DUC 2004 asked participants to create very short summaries of single documents. Documents from Task 1 came from the AP and New York Times newswires; documents for Task 3 were selected from the collection of Arabic *Agence France Presse* documents which had been translated into English by both human translators and automatic MT systems. Unlike similar tasks in previous evaluations (which placed a 10 word limit on the length of these summaries), very short summaries in DUC 2004 were limited to 75 bytes in total.

Our DUC 2004 system closely followed our DUC 2003 system (described in detail in (Lacatusu et al., 2003)) in assuming that accurate very short single-document summaries could be generated based only on the text contained in the first sentence of the document. Although most current summarization systems agree that text found in the first paragraph of a document should be considered to be most important to the generation of a single document summary ((Mani and Maybury, 1999), (Lin and Hovy, 2000)), these approaches still depend on a sentence selection algorithm in order to determine the relative usefulness of each sentence in the paragraph. In our DUC 2003 system, we eliminated the need for further sentence selection by considering only the first sentence of the document as the source of very short summaries. Although this assumption may have been too simplistic (our DUC 2003 system finished eighth overall in terms of usefulness), we believe that it performed adequately enough to be considered in DUC 2004.

We also adapted the sentence compression mechanism first applied in our DUC 2003 system to the new summary length constraints imposed in DUC 2004. Based on (Jing, 2000), our system reduced the length of sentences by making a series of decisions with regards to the right-edges of phrasal projections in a parse tree. Accounting for the length constraint in terms of bytes (as opposed to words) required no significant change in our method.

In Task 1, we submitted two sets of summaries for evaluation: (1) summaries generated from only the first sentence in the document, and (2), summaries generated from the highest ranked sentence as evaluated by the Topic Signature ranking method described in section 2. In Task 3, we followed the DUC guidelines and submitted summaries based on the automatically-translated documents prepared by ISI as well as the human-translated documents. With this latter task, the highest ranked sentence was again used to generate the summary iff any of the topic signature terms could be identified in the document; otherwise, the first sentence of the translated document was used to create the headline. The same sentence compression mechanism was applied to all summary submissions, except for the machine translated documents,

a) Washington 12-16 (AFP) - The American Defense Department (the Pentagon) reported today, Wednesday, that the bombing runs against Iraq are being especially carried out via cruise missiles launched from American aircraft carriers and B52 bombers -.	**b)** Washington 12-16 (AFP) - The ministry of defence American (Pentagon) today that bombing operations against Iraq are in particular by passing rockets fired from the aircraft carrier rocket launchers and American Me 52 -.
c) Baghdad 12- 19(AFP)- An Iraq official reported today, Saturday, that 68 Iraqi civilians were killed as a result of the American and British bombing on Iraq and that their funerals were held today in Baghdad. - The Pentagon: Bombing Being Carried out by Cruise Missiles and B52 Bombers Urgent	**d)** it was announced by the representative of Iraq in the United Nations Nizar Hamdoon on Sunday that thousands of people were killed or wounded during the four days of the bombing of Iraq.

Table 1: For the cluster d1001t: (a) Original human translated document; (b) Original machine translated document; (c) Summary of the human translated documents and (d) Summary of the machine translated documents.

where the parse tree contained too many errors and we selected the the first 75 bytes of the most important sentence to appear in the summary.

4 Summarization Focused by Biographical Questions

Question-answering (Q/A) systems identify answers to natural language questions by extracting textual passages that contain information corresponding to the question's informational goal. With trivia-like questions, similar to those evaluated in the TREC tracks, answers are returned iff a passage can be identified that meets the semantic entailments of the question's expected answer type. With complex questions (such as definition questions, list questions, or short-answer questions), the best answers may be composite ones, consisting of information extracted from multiple documents. In these cases, Q/A systems must perform a task similar to multi-document summarization and generate a coherent answer from a set of candidate answer passages.

In Task 5, participants were asked to generate short multi-document summaries that provided answers to definition questions of the form *Who is X?*, where X always corresponded to the name of a person. This kind of question-focused summarization poses a different series of challenges from the approach to topic-based summarization we presented in section 2. Since the set of documents provided for each question in this task did not necessarily focus on the named entity in the question, question-focused summarization systems must be able to identify informative passages for summarization based only on keywords found in the questions themselves. This task was complicated by the fact that the only keywords available in the questions provided by the DUC evaluators were named entities. Although the potentially restricted distribution of named entities does limit the total number of candidate passages that a summarization system needs to consider, nothing about these names themselves can be used to assist in the selection of passages to include in a summary.

We hypothesized that the most useful sentences for a definition question-focused summary would exhibit a characteristic syntactic structure. We extracted sentences from each document cluster that contained the named entity from the question either in an appositive construction (e.g. *Tutu, 68, a Nobel peace laureate for his work against apartheid in his native South Africa[...]*), or as the subject of a copula construction (e.g. *Hawking, 56, is the Lucasian Professor of Mathematics at Cambridge, a post once held by Sir Isaac Newton.*). Although we focused on only these two constructions in our DUC 2004 system, we believe that other constructions should be considered, including (1) constructions containing anaphors coreferential with the named entity, and (2) constructions involving the named entity as the subject of a Vendlerian "achievement verb" (e.g. *Robert E. Rubin [...] has taken a top position at Citigroup*) (Vendler, 1967). In future work, we expect to apply a filter for pruning erroneous and duplicate appositives, as initially implemented in (Schiffman et al., 2001).

We began by extracting all of the sentences in a document cluster that overtly mentioned the named entity specified in the question. A sentence received a score of 1.0 if it contained an appositive construction including the named entity; sentences that contained the named entity as the subject of a copula construction received a score of 0.99. If a sentence featured both an appositive and a copular verb, then it received a score of 1.99. These scores were based on an observation that appositives consistently provided more information about an entity than copular constructions. Once these scores were assigned, sentences were ranked. The top-ranked sentences were included in the summary until the length reached 700. Summaries longer than the 665 byte limit were truncated by the DUC evaluators.

This approach received the best scores in terms of coverage (as scored by human assessors using SEE). Our system also received the best scores in terms of question responsiveness (again, as determined by human assessors).

5 Results

We have focused our efforts on the multi-document summarization tasks (tasks 2, 4 and 5). In task 2 one of our runs obtained the second best score of all systems based on the ROUGE-1 scoring, which was proved to be closest to the human evaluation by (Lin and Hovy, 2003). However our priority 1 submission ranked only 8th both in the automated and the manual evaluations. The fact that two almost identical summaries have such a large difference in their rankings (ROUGE-1 scoring) can be explained by two reasons: the extra information added by the priority 3 submission is of good quality (whatever amount of information is added by the chopped sentence proved to be valuable). The second reason is that the first ranked runs have very tight scores. In fact the scores for the 8 top ranked runs fit in the confidence interval of the first one and in the confidence interval of the lowest ranked human summarizer. Even more, the confidence intervals for the first 12 systems are intersecting. The ROUGE-1 results for Task 2 are presented in Table 2.

SYSID	ROUGE-1	95% CI Lower	95% CI Upper
H	0.41828	0.40193	0.43463
F	0.41246	0.39161	0.43331
E	0.41038	0.38817	0.43259
D	0.40594	0.38700	0.42488
B	0.40428	0.37946	0.42910
A	0.39325	0.37218	0.41432
C	0.39039	0.37149	0.40929
G	0.38902	0.36793	0.41011
65	0.38224	0.36941	0.39507
104	**0.37443**	**0.36354**	**0.38532**
35	0.37430	0.36121	0.38739
19	0.37386	0.36080	0.38692
124	0.37064	0.35782	0.38346
34	0.36996	0.35613	0.38379
67	0.36971	0.35669	0.38273
102	**0.36965**	**0.35850**	**0.38080**
81	0.36844	0.35558	0.38130
45	0.36715	0.35366	0.38064
66	0.36671	0.35329	0.38013
55	0.35683	0.34527	0.36839
44	0.35006	0.33701	0.36311
56	0.35003	0.33867	0.36139
140	0.34564	0.33063	0.36065
57	0.34533	0.33297	0.35769
118	0.34381	0.32918	0.35844
138	0.34299	0.32805	0.35793
117	0.34291	0.32907	0.35675
119	0.34024	0.32366	0.35682
139	0.33959	0.32475	0.35443
95	0.33170	0.31918	0.34422
103	**0.33038**	**0.31407**	**0.34669**
93	0.32904	0.31239	0.34569
2	0.32419	0.30922	0.33916
36	0.32345	0.31172	0.33518
121	0.32051	0.30435	0.33667
120	0.30420	0.28400	0.32440
27	0.30355	0.28573	0.32137
11	0.30291	0.28340	0.32242
28	0.29522	0.28030	0.31014
123	0.28540	0.27147	0.29933
29	0.27673	0.25895	0.29451
94	0.27524	0.25920	0.29128
111	0.24190	0.23038	0.25342

Table 2: Results for Task 2, ROUGE-1 metric.

The same distribution can be observed also for the results in Task4, where our first priority run ranked third among the machine translation runs and eleventh overall. The second priority run ranked the fourth among the human translation summaries and overall. The results for Task 4 are presented in Table 3.

SYSID	ROUGE-1	95% CI Lower	95% CI Upper
Y	0.44450	0.42298	0.46602
Z	0.43263	0.40875	0.45651
X	0.42925	0.40680	0.45170
23	0.41577	0.39333	0.43821
W	0.41188	0.38696	0.43680
84	0.41012	0.38543	0.43481
145	0.40602	0.36783	0.44421
108	**0.40059**	**0.37002**	**0.43116**
69	0.39844	0.37440	0.42248
41	0.39223	0.36910	0.41536
61	0.39125	0.37061	0.41189
144	0.38827	0.36261	0.41393
146	0.38818	0.36373	0.41263
22	0.38654	0.36352	0.40956
107	**0.38615**	**0.35548**	**0.41682**
70	0.38576	0.37034	0.40118
68	0.38156	0.36420	0.39892
40	0.37960	0.35809	0.40111
85	0.37841	0.35642	0.40040
83	0.37812	0.34957	0.40667
47	0.36269	0.31643	0.40895
60	0.36055	0.34233	0.37877
42	0.35679	0.33655	0.37703
15	0.34628	0.31291	0.37965
4	0.32939	0.30321	0.35557
150	0.31581	0.28416	0.34746
48	0.30711	0.27417	0.34005
14	0.30072	0.24207	0.35937
115	0.30037	0.28271	0.31803
46	0.29599	0.26607	0.32591
114	0.28589	0.26687	0.30491
149	0.23381	0.21750	0.25012
148	0.18853	0.17300	0.20406

Table 3: Results for Task 4, ROUGE-1 metric.

For task 5 we ranked first for both responsiveness and coverage, which confirms that the person descriptions contained in appositions and copula constructions are of good quality. The results for Task 5, based on the SEE manual evaluation, are presented in Table 4.

In terms of summary quality evaluation, our system had lower scores (even if higher ranks) for Q1 and Q2 which means that automatically produced summaries still need better cohesion among sentences and they contain extraneous information.

6 Conclusions and Future Work

In this paper, we showed that work in automatic summarization can benefit from conceptually simple techniques. Lite-GISTexter produced the top-ranked summaries (in terms of coverage and responsiveness) for Task 5 by extracting information contained in a small number of syntactic constructions. In task 2, our system received the second-highest score (as scored by ROUGE) by modeling complex conceptual relations on a metric derived from the relative frequency of terms within a rel-

SYS	Q1	Q2	Q3	Q4	Q5	Q6	Q7	FUPU	PU	MPU	UPU	MU	Mean	Median	Sampl.	Responsiveness
5	1.62	2.20	1.44	1.44	1.10	1.40	1.82	0.67	5.44	2.64	2.80	11.20	0.19	0.06	0.28	1.52
16	3.12	2.70	1.58	2.04	1.70	1.36	1.46	0.43	3.04	2.16	0.88	11.20	0.16	0.01	0.26	1.30
24	3.86	3.40	1.44	2.98	1.30	1.88	1.88	0.51	4.72	2.64	2.08	11.20	0.21	0.07	0.31	1.18
30	2.90	2.42	1.38	2.00	1.46	1.30	1.30	0.39	3.06	2.34	0.72	11.20	0.20	0.05	0.30	1.42
43	3.78	2.94	1.30	2.84	1.46	1.76	1.52	0.61	5.02	3.24	1.78	11.20	0.20	0.08	0.27	1.30
49	3.34	2.88	1.54	2.12	1.62	1.78	1.36	0.68	4.74	3.10	1.64	11.20	0.21	0.05	0.30	1.40
62	3.62	3.12	1.98	1.90	2.14	1.84	1.56	0.42	4.52	3.08	1.44	11.20	0.20	0.05	0.30	1.64
71	3.32	2.96	1.84	1.44	2.42	1.64	1.66	0.40	4.14	3.06	1.08	11.20	0.21	0.05	0.32	1.50
86	3.86	3.22	1.16	2.96	1.18	2.22	1.68	0.49	4.56	2.46	2.10	11.20	0.14	0.02	0.23	1.06
96	3.30	3.06	1.60	1.82	1.66	1.68	3.54	0.46	4.02	2.70	1.32	11.20	0.22	0.07	0.32	1.44
109	**3.14**	**2.76**	**1.86**	**1.74**	**1.96**	**1.42**	**1.54**	**0.43**	**3.78**	**2.86**	**0.92**	**11.20**	**0.24**	**0.10**	**0.33**	**1.76**
116	4.52	4.04	1.48	3.40	1.76	2.14	2.14	0.68	9.02	3.72	5.30	11.20	0.17	0.04	0.27	1.00
122	2.94	2.32	1.68	1.24	2.08	1.22	1.36	0.25	2.68	2.24	0.44	11.20	0.18	0.04	0.30	1.26
125	3.52	3.22	1.62	2.12	1.94	2.76	2.18	0.39	2.98	2.06	0.92	11.20	0.19	0.04	0.30	1.40
147	3.18	2.64	1.84	1.92	1.80	1.24	1.40	0.39	3.36	2.58	0.78	11.20	0.22	0.07	0.30	1.54
A	1.00	1.06	1.00	1.00	1.00	1.11	1.06	0.50	5.61	4.94	0.67	12.06	0.48	0.53	0.42	3.60
B	1.06	1.06	1.00	1.06	1.00	1.06	1.00	0.83	7.22	5.50	1.72	10.94	0.55	0.60	0.42	3.92
C	1.24	1.36	1.28	1.04	1.16	1.40	1.12	0.64	5.12	3.72	1.40	11.36	0.35	0.25	0.38	2.84
D	1.11	1.22	1.00	1.06	1.17	1.33	1.00	0.66	6.39	4.83	1.56	10.39	0.47	0.46	0.40	3.60
E	1.39	1.33	1.00	1.11	1.00	1.33	1.22	0.67	6.17	4.50	1.67	11.11	0.47	0.41	0.41	3.32
F	1.24	1.24	1.06	1.00	1.06	1.12	1.18	0.64	6.71	5.12	1.59	11.12	0.52	0.58	0.41	3.64
G	1.17	1.00	1.00	1.00	1.00	1.28	1.06	0.82	7.22	5.44	1.78	10.61	0.54	0.54	0.40	3.48
H	1.17	1.22	1.11	1.00	1.06	1.06	1.00	0.53	5.44	4.39	1.06	11.94	0.44	0.36	0.42	3.52

Table 4: Results for Task5 - human evaluation.

evant set of documents. Although we believe that the best summarization systems of the future will depend on information obtained from a thorough understanding of discourse structure, we feel that summarization systems today can also benefit from approaches that are based on linguistically-grounded observations about texts themselves.

In future work, we plan to use coreference resolution to improve both the coverage and coherence of our summaries. We also plan to improve our redundancy detection in order to improve the qualitative scores our summaries receive. One simple way to do this would be to replace the repeat mention of entities in our short summaries with pronouns.

We also recognize that we need to employ better sentence compression techniques. Currently, our summaries contain too much nonrelevant information. If we could find reliable ways to replace these irrelevant passages with more summary-worthy snippets, we could improve the overall coverage and informativeness of our short and very short summaries. Finally, a better sentence compression algorithm would afford us the extra space to expand the number of constructions we consider in a question-focused summarization task like Task 5.

References

Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference*.

Hongyan Jing. 2000. Sentence reduction for automatic text summarization. In *In Proc. of the 6th Conference on Applied Natural Language Processing, pages 310–315, 2000.*

V. Finley Lacatusu, Paul Parker, and Sanda M. Harabagiu. 2003. Lite-GISTexter: Generating short summaries with minimal resources. In *HLT Workshop on Text Summarization (DUC 2003).*

Chin-Yew Lin and Eduard Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *Proceedings of the COLING 2000 Conference*.

Chin-Yew Lin and Eduard Hovy. 2002. NEATS in DUC 2002. In *HLT Workshop on Text Summarization (DUC 2002).*

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of 2003 Language Technology Conference (HLT-NAACL 2003).*

Inderjeet Mani and Mark T. Maybury. 1999. *Advances in Automatic Text Summarization*. MIT Press, July 1999.

M. F. Porter. 1980. An algorithm for suffix stripping. In *Program, 14(3):130-137, July 1980.*

Barry Schiffman, Inderjeet Mani, and Kristian J. Concepcion. 2001. Producing biographical summaries: Combining linguistic knowledge with corpus statistics. In *In Proceedings European Association for Computational Linguistics 2001.*

Zeno Vendler. 1967. *Linguistics in Philosophy*. Cornell University Press, Ithaca, NY.